AMENDMENTS TO THE SPECIFICATION

► Please replace the paragraph beginning on page 58, line 3 with the following replacement paragraph:

Cluster the genes using k_means, correlated-based clustering. Any standard statistical package may be used. This analysis uses the xcluster software created by Gavin Sherlock (~~http://genomewww.~~stanford.edu/sherloc-k/cluster.html). A large number of clusters are targeted so as to capture multiple, correlated patterns of variation across samples, and generally small numbers of genes within clusters;

► Please replace the paragraph beginning on page 58, line 9 with the following replacement paragraph:

Extract the dominant singular factor (principal component) from each of the resulting clusters. Again, any standard statistical or numerical software package may be used for this; this analysis uses the efficient, reduced singular value decomposition function ("SVD") in the Matlab software environment (~~http://www.~~mathworks.com/products/matlab).

► Please replace the paragraph beginning on page 58, line 14 with the following replacement paragraph:

In the analysis of the ER data in this disclosure, the original data was developed using Affymetrix arrays with 7129 sequences, of which 7070 were used (following removal of Affymetrix controls from the data.). The expression estimates used were log2 values of the signal intensity measures computed using the dChip software for post-processing Affymetrix output data (See Li, C. and Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc. Natl. Acad. Sci., 98, 31-36 (2001), and the software site ~~http://www.~~biostat.harvard.edu/complab/dchip/~~)~~. With a target of 500 clusters, the xcluster software implementing the correlation-based k_means clustering produced p=491 clusters. The corresponding p metagenes were then evaluated as the dominant singular factors of each of these clusters, as referenced above. See Table that provide tables detailing the 491 metagenes.

► Please replace the paragraph beginning on page, line with the following replacement paragraph:

The set of samples on these 7,030 genes are clustered using k-means correlated-based clustering. Any standard statistical package may be used for this; our analysis uses the xcluster software created by Gavin Sherlock at Stanford University (~~http://genome-www.~~stanford.edu/sherlock/- cluster.html). We defined a target of 500 clusters and the xcluster routine delivered 496 in this analysis.

► Please replace the paragraph beginning on page, line with the following replacement paragraph:

The dominant singular factor (principal component) from each of the 496 clusters is extracted. Again, any standard statistical or numerical software package may be used for this; this analysis uses the reduced singular value decomposition function (svd) in Matlab. (~~http://www.~~mathworks.com/products/matlab).

► Please replace the paragraph beginning on page 62, line 12 with the following replacement paragraph:

Hybridization procedures and parameters. The amount of starting total RNA for each reaction was 20 μg. Briefly, first strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An in vitro transcription reaction was performed to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C. for 35 min. The fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin to Affymetrix GeneChip Human U95Av2 arrays at 45°C. for 16 hr, according to the Affymetrix protocol (~~www.~~affymetrix.com and Pittman Ms-NG 21-~~www.~~affymetrix.com/products/arrays/specific/hgu95.affx). The arrays contain over 12,000 genes and ESTs. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated antistreptavidin antibody (Vector Laboratories, Burlingame, Calif.)

at 3μg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent. Each sample was hybridized once.

► Please replace the paragraph bridging pages 62-63 with the following replacement paragraph:

Measurement data and specifications. Scans were performed with an Affymetrix GeneChip scanner and the expression value for each gene was calculated using the Affymetrix Microarray Analysis Suite (v5.0), computing the expression intensities in `signal` units defined by software. Scaling factors were determined for each hybridization based on an arbitrary target intensity of 500. Scans were rejected if the scaling factor exceeded a factor of 25, resulting in only one reject. Array design. All assays employed the Affymetrix Human U95Av2 GeneChip. The characteristics of the array are detailed on the Affymetrix web site (~~www.~~affymetrix.com/products/arrays/specific/hgu95.affk).

► Please replace the paragraph beginning on page 64, line 14 with the following replacement paragraph:

The set of samples on these 7,030 genes are clustered using k-means correlated-based clustering. Any standard statistical package may be used for this; our analysis uses the xcluster software created by Gavin Sherlock at Stanford University (~~http://genome-www.~~stanford.edu/sherlock/- cluster.html). We defined a target of 500 clusters and the xcluster routine delivered 496 in this analysis

► Please replace the paragraph beginning on page 64, line 20 with the following replacement paragraph:

The dominant singular factor (principal component) from each of the 496 clusters is extracted. Again, any standard statistical or numerical software package may be used for this; this analysis uses the reduced singular value decomposition function (svd) in Matlab. (~~http://www.~~mathworks.com/products/matlab).

► Please replace the paragraph bridging pages 79-80 with the following replacement paragraph:

Hybridization procedures and parameters. The amount of starting total RNA for each reaction was 20 μg. Briefly, first strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An in vitro transcription reaction was performed to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C. for 35 min. The fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin to Affymetrix GeneChip Human U95Av2 arrays at 45°C for 16 hr, according to the Affymetrix protocol (www.affymetrix.com and Pittman Ms-NG 21 www.affymetrix.com/products/arrays/specific/hgu95.affx). The arrays contain over 12,000 genes and ESTs. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated antistreptavidin antibody (Vector Laboratories, Burlingame, Calif.) at 3 μg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent. Each sample was hybridized once.

► Please replace the paragraph beginning on page 80, line 23 with the following replacement paragraph:

Array design. All assays employed the Affymetrix Human U95Av2 GeneChip. The characteristics of the array are detailed on the Affymetrix web site (~~www.~~affymetrix.com/products/arrays/specific/hgu95.affk).

► Replace the Abstract starting on page 98, line 3 with the following replacement paragraph:

The statistical analysis described and claimed is a predictive statistical tree model that overcomes several problems observed in prior statistical models and regression analyses, while ensuring greater accuracy and predictive capabilities. Although the claimed use of the predictive statistical tree model described herein is directed to the prediction of a disease in individuals, the

claimed model can be used for a variety of applications including the prediction of disease states, susceptibility of disease states or any other biological state of interest, as well as other applicable non-biological states of interest. ~~This model first screens genes to reduce noise, applies k-means correlation-based clustering targeting a large number of clusters, and then uses singular value decompositions (SVD) to extract the single dominant factor (principal component) from each cluster. This generates a statistically significant number of cluster-derived singular factors, that we refer to as metagenes, that characterize multiple patterns of expression of the genes across samples. The strategy aims to extract multiple such patterns while reducing dimension and smoothing out gene-specific noise through the aggregation within clusters. Formal predictive analysis then uses these metagenes in a Bayesian classification tree analysis. This generates multiple recursive partitions of the sample into subgroups (the "leaves" of the classification tree), and associates Bayesian predictive probabilities of outcomes with each subgroup. Overall predictions for an individual sample are then generated by averaging predictions, with appropriate weights, across many such tree models. The model includes the use of iterative out-of-sample, cross-validation predictions leaving each sample out of the data set one at a time, refitting the model from the remaining samples and using it to predict the hold-out case. This rigorously tests the predictive value of a model and mirrors the real-world prognostic context where prediction of new cases as they arise is the major goal.~~